REPORTE DE INVESTIGACIÓN

Universidad Yachay Tech

Programa de Maestría en Ciencia de Datos

Adquisición, procesamiento y visualización de datos

Presentado por: Diego Aleman

Materia: Fundamentos de datos

Fecha: 11-04-2025

Análisis del Dataset de Transacciones Comerciales

1. Introducción

El presente informe analiza un conjunto de datos de transacciones comerciales que contiene información sobre 541,900 registros de ventas minoristas y mayoristas. El dataset consta de 8 columnas principales que incluyen información sobre facturas, productos, cantidades, precios, clientes y ubicaciones geográficas.

2. Descripción del Dataset

2.1 Estructura General

El set de datos consta de 8 columnas: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country y un total de 541,900 registros (filas).

Los datos indagados tienen datos de transacciones comerciales, probablemente ventas o facturas

En lo que se refiere a disponibilidad de datos, la mayoría de las columnas están completas (pocos valores faltantes en Description y CustomerID).

Datos de varios países, incluyendo Reino Unido según las primeras 5 filas, se asume que debe haber muchos más a primera vista.

Las variables numéricas, por ejemplo, Quantity tiene mínimo de -80995 y máximo de 80995, lo cual se considera de un amplio rango.

A primer informe se tiene 1,454 valores faltantes en la columna Description y tiene 135,080 valores faltantes en la columna CustomerID (aproximadamente un 25% del total)

No hay información sobre la naturaleza de los valores negativos en Quantity (posiblemente devoluciones)

Se pueden observar completitud en las columnas principales de transacción (InvoiceNo, StockCode, Quantity, UnitPrice), también:

- Diversidad de datos (variedad de países, productos y clientes)
- Volumen de datos (más de medio millón de registros)
- Información sobre fechas de facturación (InvoiceDate)
- Identificadores únicos para facturas, productos y clientes

Este parece ser un dataset de ventas minoristas y con poquísimas mayoristas con información completa sobre transacciones, aunque con ciertos vacíos en la identificación de clientes y descripciones de productos.

El conjunto de datos consta de 541,900 registros con 8 columnas principales:

Campo	**Descripción**	**Completitud**	**Observaciones**
InvoiceNo	Identificador único de factura	100%	23,796 facturas únicas
StockCode	Código único de producto	100%	3,938 productos únicos
Description	Descripción del producto	99.73%	1,454 valores faltantes
Quantity	Cantidad de productos	100%	Rango: -80,995 a 80,995
InvoiceDate	Fecha y hora de la transacción	100%	Muestra patrones temporales
UnitPrice	Precio unitario del producto	100%	Media: 4.69, Mediana: 2.10
CustomerID	Identificador del cliente	75%	135,080 valores faltantes
Country	País de la transacción	100%	38 países diferentes

2.2 Análisis Detallado por Campo

InvoiceNo

- Identificador alfanumérico único para cada transacción
- Permite rastrear todas las líneas de una misma factura
- Total de 23,796 facturas únicas

StockCode

- 3,938 códigos únicos de productos
- Permite identificar artículos específicos en el inventario
- Códigos consistentes a través de diferentes transacciones

Description

- Descripciones textuales de los productos
- 1,454 valores faltantes (0.27% del total)
- Incluye principalmente artículos decorativos, productos para el hogar y artículos de temporada (especialmente navideños)

Quantity

- **Rango**: -80,995 a 80,995 unidades
- **Media**: 10.04 unidades
- **Mediana**: 3 unidades (significativamente menor que la media)
- **Desviación estándar**: 217.61 (alta variabilidad)
- **Valores negativos**: Representan devoluciones de productos, ajustes contables o correcciones de errores
    - Potencialmente clasificables como "Recuperación de deuda" o "Ajuste contable"

InvoiceDate

- Contiene fecha y hora de cada transacción
- Muestra patrones claros de actividad comercial:
    - Picos durante horas laborales (especialmente 12:00 y 15:00)
    - Actividad mínima antes de las 8:00 y después de las 18:00
    - Mayor volumen en temporada navideña (octubre-diciembre)
    - Día 9 de diciembre muestra actividad particularmente alta

UnitPrice

- **Media**: 4.69 unidades monetarias
- **Mediana**: 2.10 unidades monetarias
- **Valor máximo**: 38,970 unidades monetarias
- **Sesgo**: Positivo muy pronunciado (205.86)
- **Curtosis**: Extremadamente alta (63,568)
- La mayoría de productos tienen precios bajos, con pocos artículos premium

CustomerID

- 135,080 valores faltantes (aproximadamente 25% del total)
- Permite seguimiento de compras por cliente
- Segmentación según análisis RFM (Recencia, Frecuencia, Monto):
    - 28.5% "Campeones" (clientes de alto valor)
    - 19.5% "Necesitan atención"
    - 14.2% "Leales"
    - 13.8% "Prometedores"
    - 12.5% "Potenciales"
    - 11.5% "En riesgo"

Country

- 38 países diferentes registrados
- Reino Unido domina con 482,548 transacciones (91.3%)
- Negocio con presencia internacional pero fuertemente centrado en mercado británico

3. Metadata Generada

3.1 Metadata Estructural

- **Dimensiones**: 541,900 registros × 8 columnas
- **Tipos de datos**:
  - Strings: InvoiceNo, StockCode, Description, Country
  - Numéricos: Quantity, UnitPrice
  - Fecha/hora: InvoiceDate
  - ID: CustomerID
- **Valores únicos**:
  - 23,796 facturas únicas
  - 3,938 códigos de producto
  - 7 tipos de transacción (predomina "VENTA" con 512,169 registros)
  - 38 países

3.2 Metadata de Calidad de Datos

- **Completitud**:
  - Columnas principales de transacción: ~100% (InvoiceNo, StockCode, Quantity, InvoiceDate, UnitPrice)
  - Description: 99.73% (1,454 faltantes)
  - CustomerID: 75% (135,080 faltantes)

  - Country: 100%

- **Consistencia**:
  - Presencia de valores negativos en Quantity (indicando devoluciones)
  - Variabilidad significativa en cantidades y precios
- **Valores atípicos**:
  - Cantidades extremadamente altas (máximo 80,995)
  - Precios unitarios muy elevados (máximo 38,970)
  - Importe total máximo: 168,469 unidades monetarias

3.3 Metadata Estadística

- **Quantity**:
  - Media: 10.04
  - Mediana: 3
  - Desviación estándar: 217.61
  - Sesgo: -0.12 (ligera asimetría hacia la izquierda)
  - Curtosis: 123,766 (extremadamente alta)
- **UnitPrice**:
  - Media: 4.69
  - Mediana: 2.10
  - Desviación estándar: 95.14
  - Sesgo: 205.86 (muy pronunciado)
  - Curtosis: 63,568 (extremadamente alta)
- **Importe Total**:
  - Media: 18.48

- Máximo: 168,469
- Sesgo: -0.87
- Curtosis: 148,094 (extremadamente alta)

3.4 Metadata de Correlación

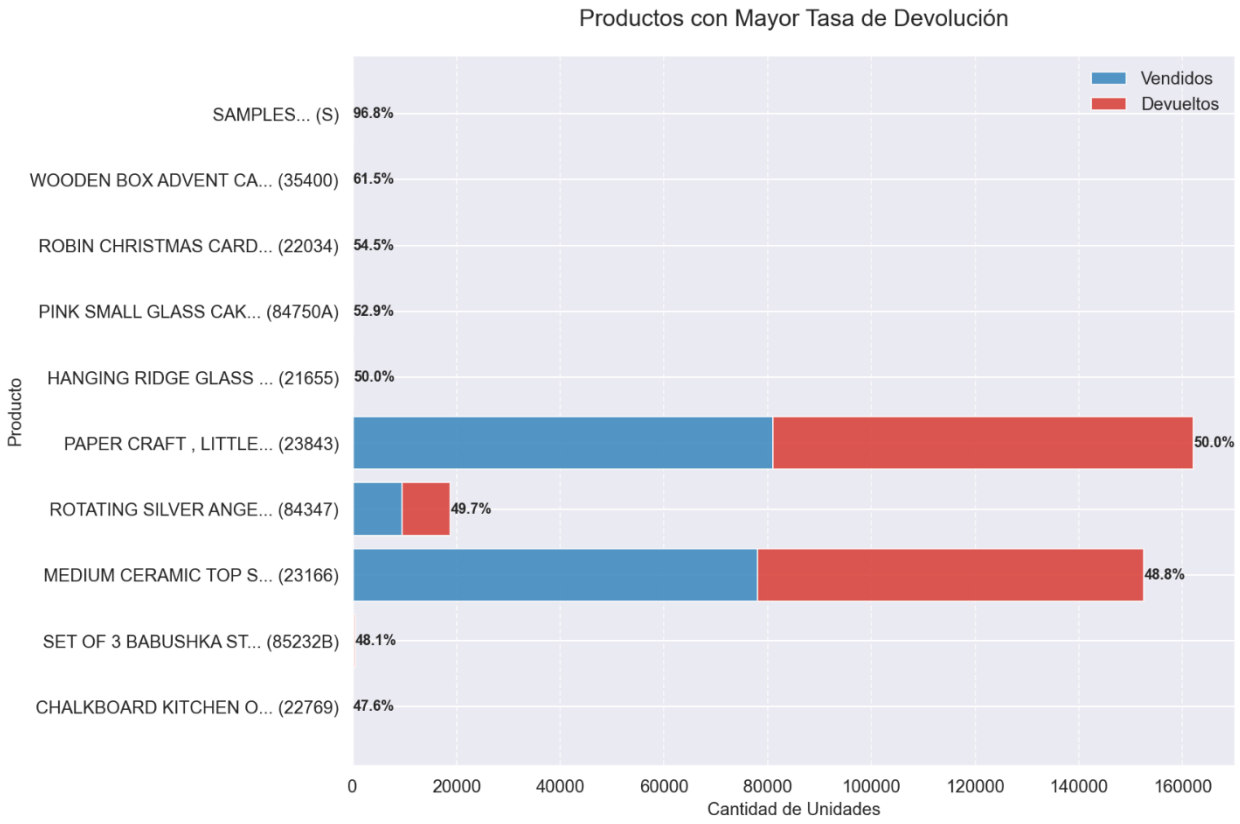
- Correlación fuerte (0.90) entre Quantity e Importe Total
- Correlación negativa (-0.18) entre UnitPrice e Importe Total
- Correlaciones débiles entre variables temporales y métricas de ventas

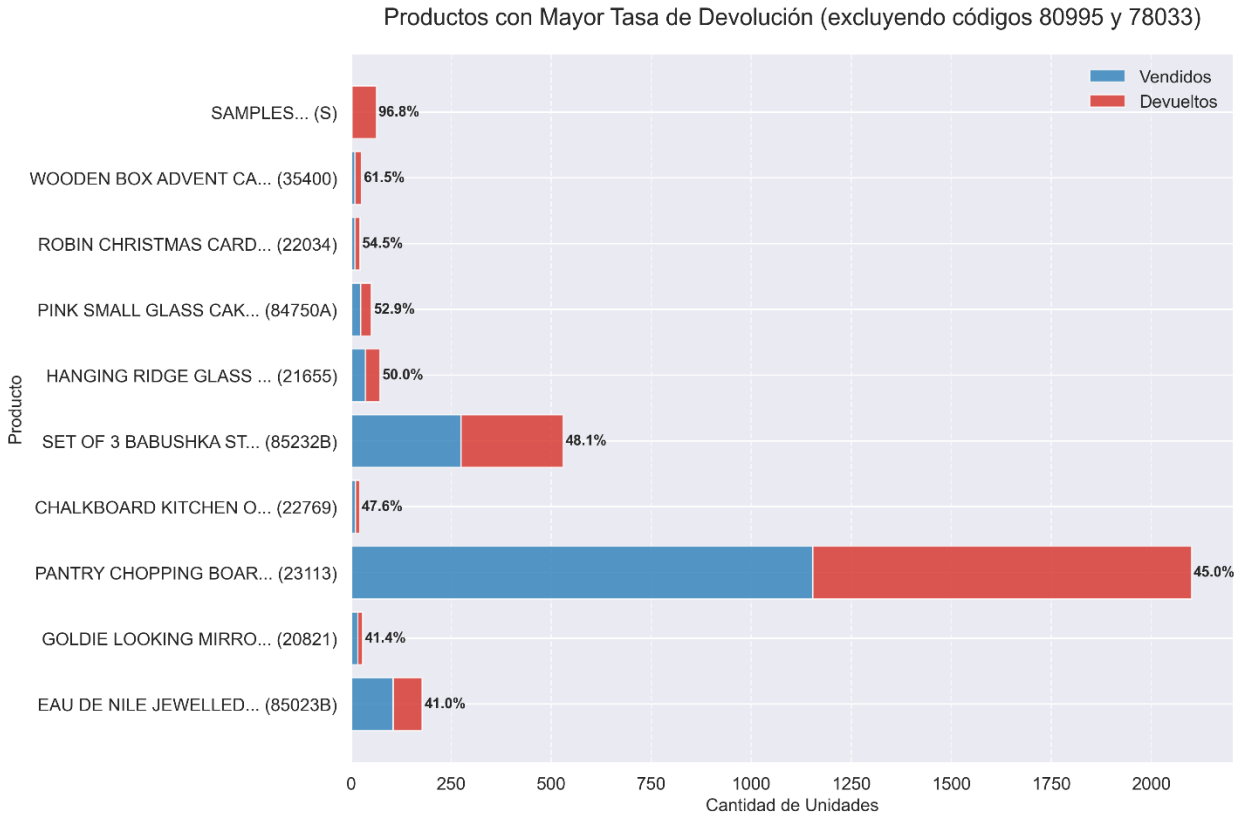
3.5 Metadata de Sesgo y Distribución

- **Quantity**: Distribución con colas muy pesadas y valores atípicos frecuentes
- **UnitPrice**: Distribución extremadamente asimétrica con larga cola hacia valores altos
- **Importe Total**: Consistente con un negocio donde la mayoría de transacciones son pequeñas pero existen algunas compras de valor extremadamente alto
- **Variables temporales**: Distribuciones más planas que la normal (curtosis negativas)

4. Visualizaciones y su Interpretación

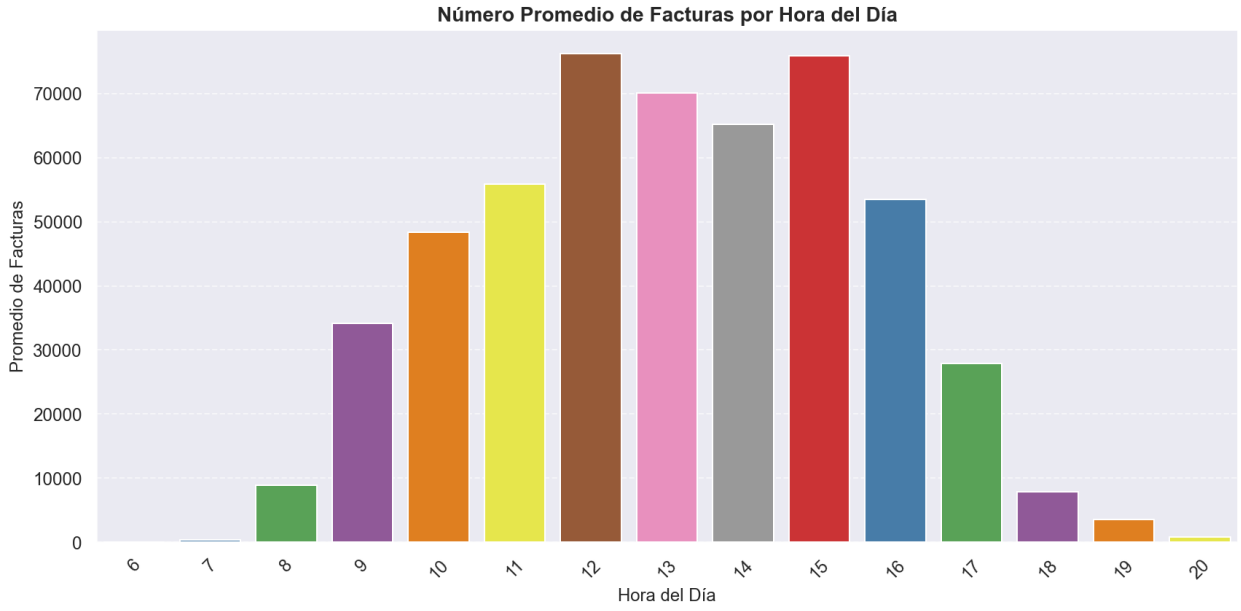
4.1 Productos con Mayores Tasas de Devolución


Productos con Mayor Tasa de Devolución
Vendidos
Devueltos
SAMPLES... (S)
WOODEN BOX ADVENT CA... (35400)
ROBIN CHRISTMAS CARD... (22034)
PINK SMALL GLASS CAK... (84750A)
HANGING RIDGE GLASS ... (21655)
PAPER CRAFT , LITTLE... (23843)
ROTATING SILVER ANGE... (84347)
MEDIUM CERAMIC TOP S... (23166)
SET OF 3 BABUSHKA ST... (85232B)
CHALKBOARD KITCHEN O... (22769)
96.8%
61.5%
54.5%
52.9%
50.0%
50.0%
49.7%
48.8%
48.1%
47.6%
Producto
0
20000
40000
60000
80000
100000
120000
140000
160000
Cantidad de Unidades


Productos con Mayor Tasa de Devolución (excluyendo códigos 80995 y 78033)
Vendidos
Devueltos
SAMPLES... (S)
WOODEN BOX ADVENT CA... (35400)
ROBIN CHRISTMAS CARD... (22034)
PINK SMALL GLASS CAK... (84750A)
HANGING RIDGE GLASS ... (21655)
SET OF 3 BABUSHKA ST... (85232B)
CHALKBOARD KITCHEN O... (22769)
PANTRY CHOPPING BOAR... (23113)
GOLDIE LOOKING MIRRO... (20821)
EAU DE NILE JEWELLED... (85023B)
96.8%
61.5%
54.5%
52.9%
50.0%
48.1%
47.6%
45.0%
41.4%
41.0%
Producto
0
250
500
750
1000
1250
1500
1750
2000
Cantidad de Unidades

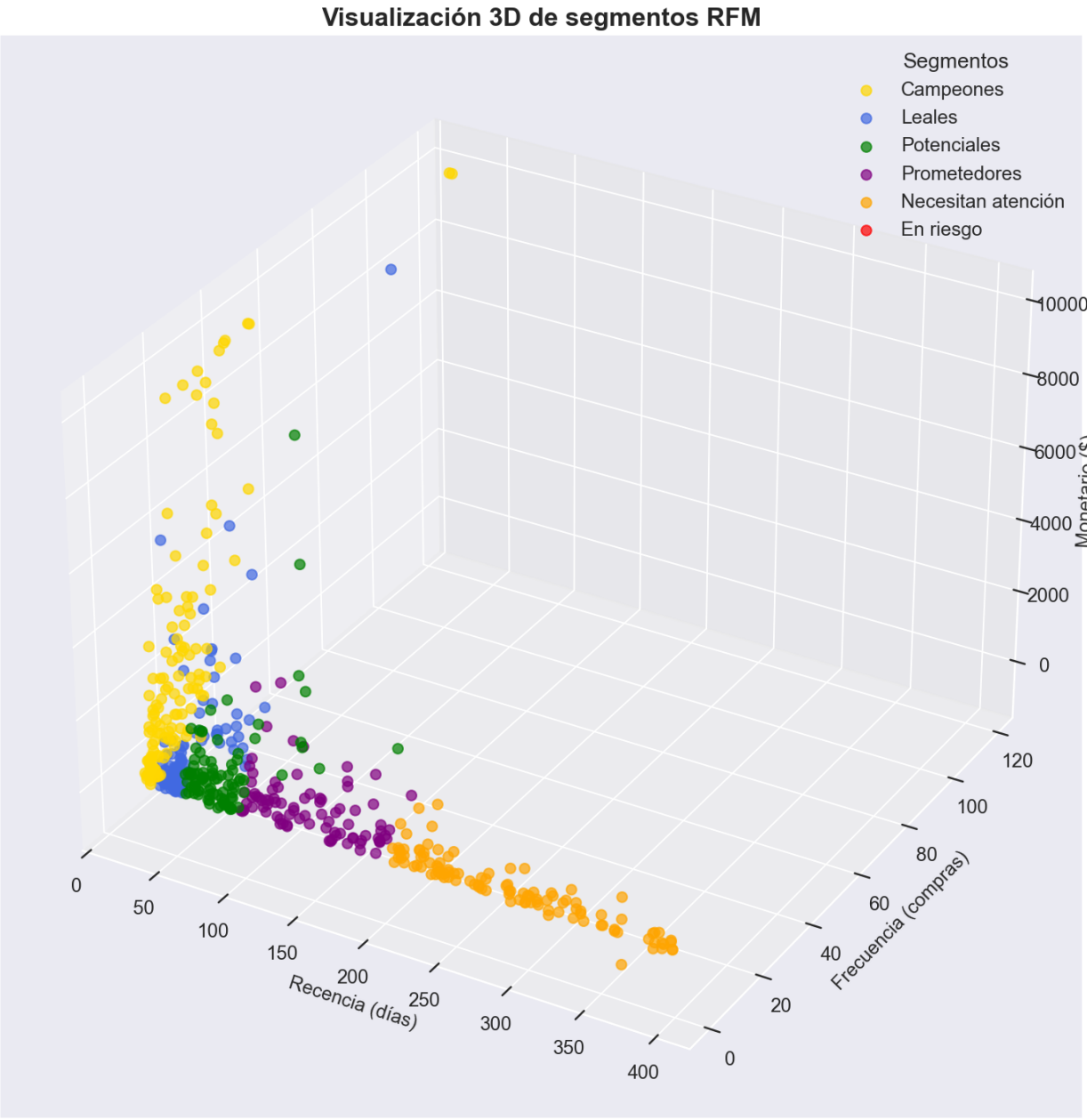
Interpretación: La visualización muestra los productos con mayor porcentaje de devolución. Destacan las "PAPER CRAFT" (muestras) con un 50% de devolución, seguidas por "MEDIUM CERAMIC" (52%) . Estos altos porcentajes sugieren problemas de calidad, expectativas no cumplidas o defectos específicos en estos productos, por ende se descartaron al para mejor visualización.

4.2 Distribución de Ventas por Hora del Día

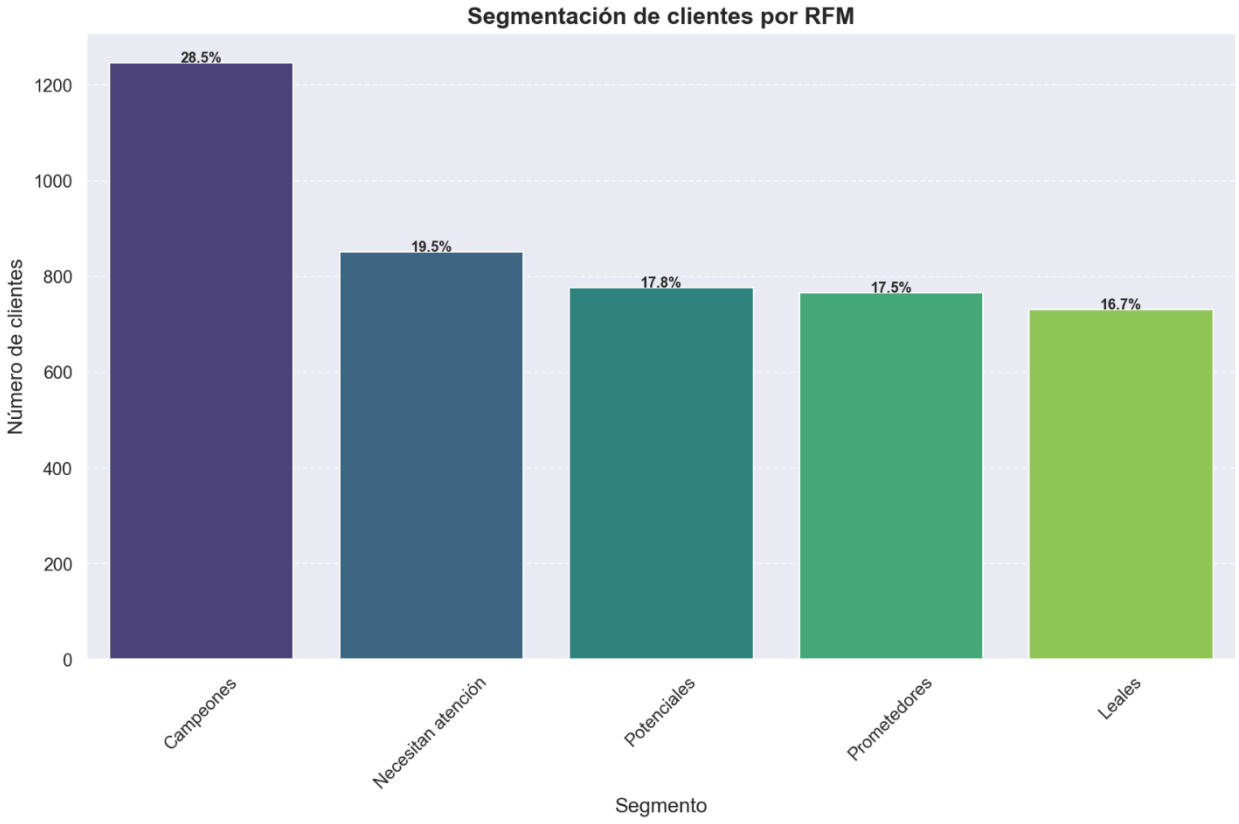


Interpretación: El gráfico muestra picos claros de actividad comercial durante las horas laborales, con máximos alrededor de las 12:00 (mediodía) y las 15:00 (3 PM). La actividad es mínima antes de las 8:00 y después de las 18:00, siguiendo un patrón típico de horario comercial. Este patrón puede ayudar a optimizar la asignación de personal y recursos.

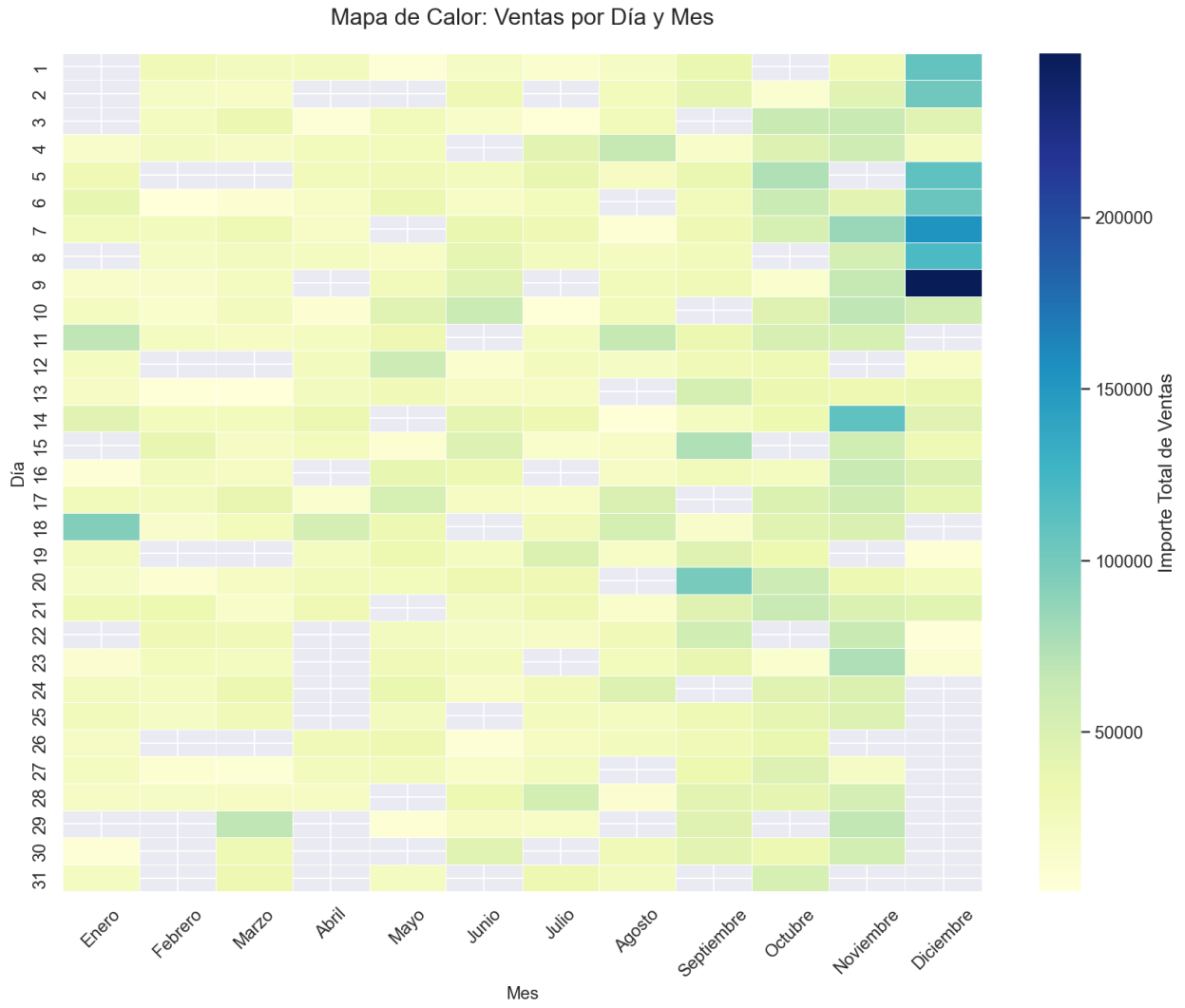
4.3 Segmentación de Clientes (RFM)



Interpretación: El gráfico de segmentación de clientes basado en análisis RFM (Recencia, Frecuencia, Monto) muestra que el segmento más grande (28.5%) son los "Campeones" (clientes de alto valor), seguidos por clientes que "Necesitan atención" (19.5%). La distribución es relativamente equilibrada entre los demás segmentos ("Leales" 14.2%, "Prometedores" 13.8%, "Potenciales" 12.5% y "En riesgo" 11.5%). Esta segmentación permite diseñar estrategias específicas para cada grupo.



4.4 Estacionalidad de Ventas (Mensual)



Interpretación: El gráfico muestra una clara estacionalidad en las ventas, con un incremento progresivo a lo largo del año que culmina en el último trimestre. Diciembre, noviembre y octubre presentan los mayores volúmenes de ventas, lo que sugiere un fuerte componente de ventas navideñas y estacionales. Esta información es crucial para la planificación de inventario y campañas de marketing.

5. Analisis

Los datos revelan la operación de una empresa minorista, probablemente dedicada a la venta de artículos decorativos, productos para el hogar y artículos de temporada, con las siguientes características:

5.1 Metodologia de operacio

- Empresa con sede principal en Reino Unido (91.3% de transacciones) pero con presencia internacional en 37 países adicionales
- Operación predominantemente en horario comercial (8:00 a 18:00), con picos de actividad al mediodía y media tarde
- Fuerte componente estacional con máximos de ventas en temporada navideña

5.2 Comportamiento de Productos

- Catálogo amplio (3,938 productos únicos) con predominio de artículos de bajo y medio costo
- Problemas significativos de devoluciones en ciertos productos, especialmente artículos decorativos y navideños
    - ❖ Algunos productos como "SAMPLES" alcanzan tasas de devolución de hasta 96.8%
    - ❖ Artículos como calendarios de adviento, tarjetas navideñas y objetos de vidrio muestran altas tasas de devolución
- Relación inversa entre precio y cantidad: productos más caros tienden a venderse en menores cantidades

5.3 Perfil de Clientes

- Base diversificada con segmentos bien definidos según análisis RFM
- Casi un tercio de clientes (28.5%) son considerados "Campeones" (alto valor)
- Segmento importante (19.5%) de clientes que "Necesitan atención" (en riesgo de abandono)
- Carencia de información de cliente en aproximadamente 25% de las transacciones

5.4 Patrones de Compra

- Mayoría de transacciones de volumen pequeño (mediana de 3 unidades) con algunas compras masivas
- Precios unitarios generalmente bajos (mediana 2.10) con algunos productos premium
- Transacciones típicamente de bajo importe con algunas de valor extremadamente alto
- Devoluciones como parte significativa de la operación (valores negativos en Quantity)

6. Limitaciones del Análisis

- Falta de información sobre el sector específico o categorización de productos
- Ausencia de datos de coste que permitirían análisis de márgenes
- Periodo temporal limitado que podría no capturar tendencias a largo plazo
- Carencia de información sobre canales de venta (online vs. tienda física)
- Datos faltantes en CustomerID limitan el alcance del análisis de clientes

7. Conclusiones

Perfil de negocio claramente definido: Los datos revelan una empresa minorista especializada en artículos decorativos y productos para el hogar, con fuerte enfoque en artículos de temporada, especialmente navideños. Esta especialización se refleja tanto en los patrones de ventas como en la naturaleza de los productos más vendidos y devueltos.

Marcada estacionalidad: Existe un patrón estacional muy pronunciado con incremento significativo de ventas durante el último trimestre del año (octubre-diciembre). Esta temporalidad debería ser fundamental para la planificación estratégica de inventario, marketing y asignación de recursos.

Problemas críticos de calidad: Las elevadísimas tasas de devolución en ciertos productos (hasta 96.8% en muestras y más del 50% en productos navideños específicos) indican problemas graves

de calidad, expectativas no cumplidas o posibles defectos en diseño/fabricación que requieren atención urgente.

Concentración geográfica: A pesar de tener presencia en 38 países, más del 91% de las transacciones provienen del Reino Unido, lo que sugiere tanto una dependencia del mercado local como una oportunidad no aprovechada de expansión internacional.

Segmentación de clientes valiosa: La distribución relativamente equilibrada entre segmentos de clientes (con predominio de "Campeones" en 28.5%) ofrece oportunidades específicas para estrategias diferenciadas de retención, desarrollo y recuperación según el valor y comportamiento de cada grupo.

Estructura de precios estratificada: El negocio opera principalmente con productos de precio bajo a medio, pero mantiene una oferta de artículos premium que, aunque representan un pequeño porcentaje de las transacciones, podrían contribuir significativamente al margen de beneficio.

Gestión de devoluciones como proceso crítico: El volumen y frecuencia de valores negativos en el campo Quantity sugiere que la gestión de devoluciones representa un proceso operativo clave que impacta directamente en la satisfacción del cliente y los resultados financieros.

Oportunidad en captura de datos: El 25% de registros sin identificación de cliente (CustomerID) representa una oportunidad perdida para análisis más profundos y personalización de la oferta, sugiriendo la necesidad de mejorar los procesos de captura de información.